Analysis of Biological Networks:
Protein-Protein Interaction Data Preprocessing

Lecturer: Roded Sharan Scribe: Nadav Sofy and Tomer Levi [*]

Lecture 7, December 7, 2006

1 Noise in PPI data

1.1 Protein Interaction Detection Methods

In Lecture 1 we learned about two methods for identifying protein interactions: Yeast Two-Hybrid and CoImmunoprecipitation. These techniques are designed to identify physical bindings between proteins.

1.1.1 Yeast Two-Hybrid (Y2H)

The yeast two-hybrid technique [17, 9] allows the detection pair-wise protein interactions. It exploits the modular property typical of many eukaryotic transcription factors, which can be usually decomposed in two distinct modules, one directly binding to DNA (DB, DNA-binding domain) and the other activating transcription (AD, transcriptional activating domain). The first component, DB, is able to bind to DNA even by itself, while the second module, AD, will activate transcription only if physically associated to a binding domain. In the two-hybrid experiment the test proteins are expressed as fusion proteins (hybrids) with a DNA-binding domain (DB, the bait) and a transcriptional activating domain (AD, the prey). Fusions partners are coexpressed in yeast nucleus where a protein-protein interaction is identified thanks to the activation of the reporter gene, which can be detected and measured.

Figure 1 shows that the two proteins whose interaction is under scrutiny, here indicated as bait and prey, are expressed as fusion proteins, respectively, with a binding domain (BD) and an activation domain (AD). If an interaction between bait and prey takes place, the complex formed activates the transcription of the reporter gene, allowing, as a consequence, the detection of the interaction itself.

Advantages:

1. Y2H is an in vivo technique, take place inside the organism (in the nucleus). All conditions are natural, there are no artificial lysis or washing steps.
2. This technique detects even transient and unstable interactions.
3. It is independent of endogenous protein expression.
4. It has fine resolution, enabling interaction mapping within proteins.
5. It does not require any previous knowledge of the proteins to be tested and can be performed once the corresponding genes are known, thus being suitable for large-scale applications.

[*] Based on a scribe by Michal Erel and Eli Haham

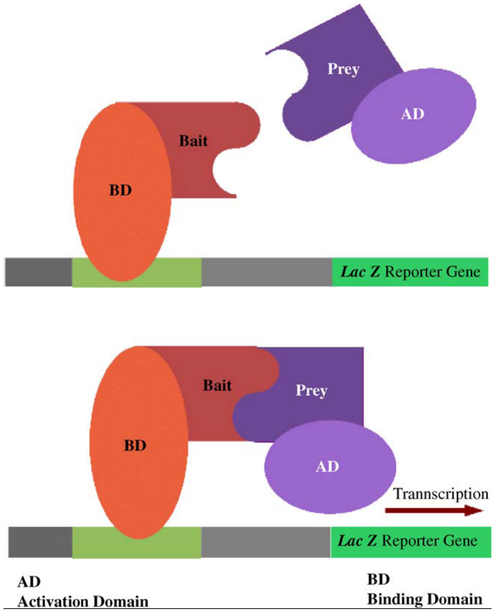


Figure 1: Source [3]: Illustration of the Y2H technique .

Drawbacks:

1. It only detects binary interactions and does not identify cooperative binding. We do not know which of the detected interactions are co-occurring (”static picture”).
2. It takes place in the nucleus, so many proteins are not in their native compartment. This could lead to both false-positives and false-negatives. (i) False positives: Interactions that are falsely identified in the experiment. The reason is that even if two proteins potentially interact into the nucleus, where this techniques takes place, it could happen that they never find close to each other because they could be localized in different cell types or could be expressed at different points. (ii) False-negatives: Interactions that where unidentified in the experiment. If two proteins interact outside the nucleus, they may not do so inside the nucleus and Y2H will not detect their interactions.
3. Some kinds of proteins, such as transcription factors, cannot be studied with this technique since their hybrids could activate the transcription even in absence of any interaction.
4. The extensive use of artificially made hybrids could lead to conformational changes in the bait and prey proteins thus preventing transcriptional activation. This is one of the possible causes of false negative interactions .
5. It predicts possible interactions, but is unrelated to the physiological setting.

1.1.2 Coimmunoprecipitation (coIP)

After the development of ultra-sensitive mass spectrometric techniques for protein identification, new experimental procedures, besides two-hybrid screens, have been used to produce large-scale results for protein-protein interactions, such as purification of protein complexes. This procedure is made up of three main steps:

1. Isolation of the bait or target protein.
2. Affinity purification of the complex.
3. Identification by mass spectrometry of proteins belonging to the complex.

The protein of interest is isolated and fused to an affinity tag, by using one of the two protocols:

1. Tandem affinity purification (TAP) [17, 13].
2. High-throughput mass spectrometric protein complex identification (HMS-PCI) [6].

TAP consists of two successive affinity purifications, using two tags fused with the bait and leading to the isolation of the target protein together with its associated proteins. In HMS-PCI, first, the bait proteins contain flag epitope tag and are overexpressed from GAL1 to tet promoters, and afterwards, the protein assemblies are isolated in one-step immunoaffinity purification followed by resolution on SDS-PAGE, digestion and MS, and MS/MS analysis. Unfortunately, comparison of results obtained through complex purification with yeast two-hybrid data shows a very small overlap. A possible explanation could rely on the fact that cooperative binding embodied by complexes is not only the result of a sum of pair-wise interactions. Indeed, the main difference between complex purification methods and two-hybrid system relies in the identification of whole complexes isolated in a single step, thus detecting cooperative interactions between proteins which cannot result from two-hybrid screens, where the strategy adopted is based on the bi-modular properties of transcription factors.

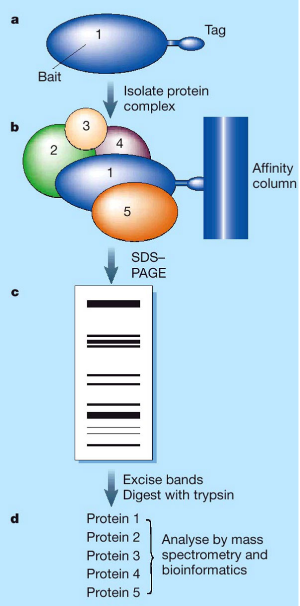


Figure 2: The coIP process. Process stages: a) Specific protein baits are prepared. b) The expressed bait proteins carry an affinity tag that allows the purification of the bait protein and the associated proteins. c) Purified protein complexes are resolved, and discrete protein bands are excited and digested into small peptide fragments. d) Peptides are identified using mass spectrometry methods. The identity of the protein associated with a given bait is determined by comparing its peptide fingerprint against databases.

Advantages:

1. Since coIP detects whole complexes of protein interactions (in contrast to Y2H that detects only pair-wise protein interactions) it is capable of detecting interactions that depend on higher order complexes.
2. CoIP detects real complexes in physiological conditions, since interactions take place in native environment.
3. CoIP is an in vivo technique that employs only one artificially made protein (the bait), instead of two as in two-hybrid procedure, thus minimizing possible changes in conformational properties that could lead to steric interference.
4. In order to test the validity of a complex identification, several components of the same complex can be used as tagged baits.

Drawbacks:

1. CoIP might miss some complexes that are not present under the given conditions.
2. Low-abundance protein might be missed.
3. Over expression of bait proteins might lead to the detection of false-positive interactions (HMS-PCI).
4. Tagging may disturb complex formation; and loosely associated components may be washed off during purification.

1.2 Reproducibility

Previous studies tested the reproducibility of different methods such as coIP and Y2H. As an example consider the Septin complex depicted in figure 3:

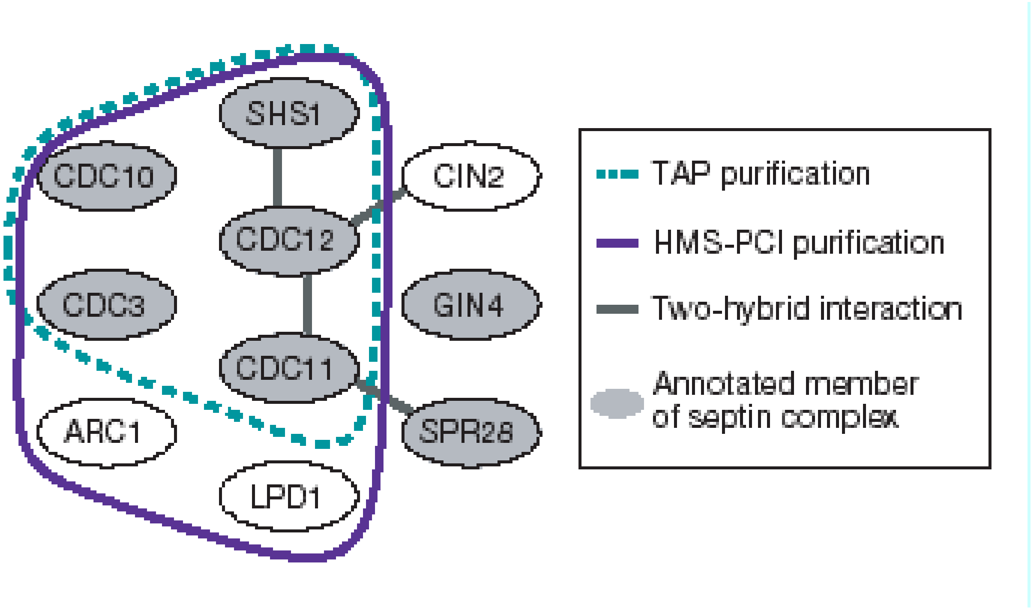


Figure 3: Source [18]: The Septin complex and its mapping by the different methods . Gray areas are the true members of a septin complex. As can be seen HMS-PCI revealed 5 out of 7 members of the complex, while TAP in addition to stating those same 5 proteins also included 2 false positives. Y2H indicates 4 interactions, 3 out of which involve the true members of the complex.

Interaction data from high throughput experiments comes in the binary interactions format (Y2H) or groups of interacting partners format (coIP). We can measure this data in different ways: counting proteins or counting binary interactions. The method we choose may effect the accuracy of the results. In the example in Figure 3, Y2H will have a false positive rate of 1/5 when counting proteins, but 1/4 when counting interactions. When binary interactions are known it seems more appropriate to work at the interaction level. When they are not know as in the case of complexes it seems more appropriate to work at the protein level.

As can be seen from Figure 4 there is a very small lap between the datasets from the different experiments. The maximal overlap is demonstrated between Ito and Uetz, and reaches 14% of the size of the smaller dataset. In some cases there is no overlap at all. There is also a very small overlap between any of these datasets and the set derived from small scale experiments (which varies from 0% to 17%).

Another example for small overlap, demonstrated in Figure 5, is the lap between Gavin spoke dataset and Ho spoke dataset as well as Ho matrix dataset which is smaller than the lap of any other dataset overlap.

There are several reasons for the lack of overlap such as: the use of different yeast strains, the use of non-physiological (artificial) conditions in the experiment, the different methods used in the experiments. We can actually

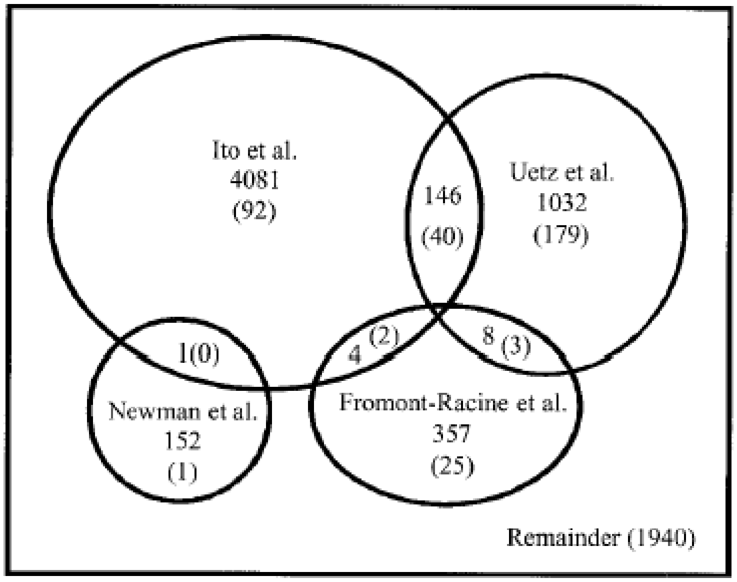


Figure 4: Source [4]: A Venn diagram illustrating the overlap between the PPI datasets in yeast. Each Oval in the figure represents a high throughput Y2H study, and the overlaps between the studies are given by the intersections. The numbers in parentheses represent the interactions that have been determined by small scale methods .

discover proteins that can physically interact, but will never interact in-vivo, due to different localization or expression at different time during the cell cycle.

1.3 Functional Profiles

An independent measure to assess the quality of experiment is the functional correlation between the interacting proteins, which is represented by a matrix whose. Each axis of the matrix represents the entire yeast genome, which has been subdivided into functional categories using MIPS. Proteins with related functions preferentially interact with each other, and therefore it is expected to see a large number of interactions along the diagonal. Interactions outside the diagonal are likely to be false positives.

Figure 6 shows an analysis of large scale protein interaction data sets (HMS-PCI and TAP) represnted by matrix model, as well as analysis of the overlap between high throughput methods and reference data of known complexes (such as MIPS).

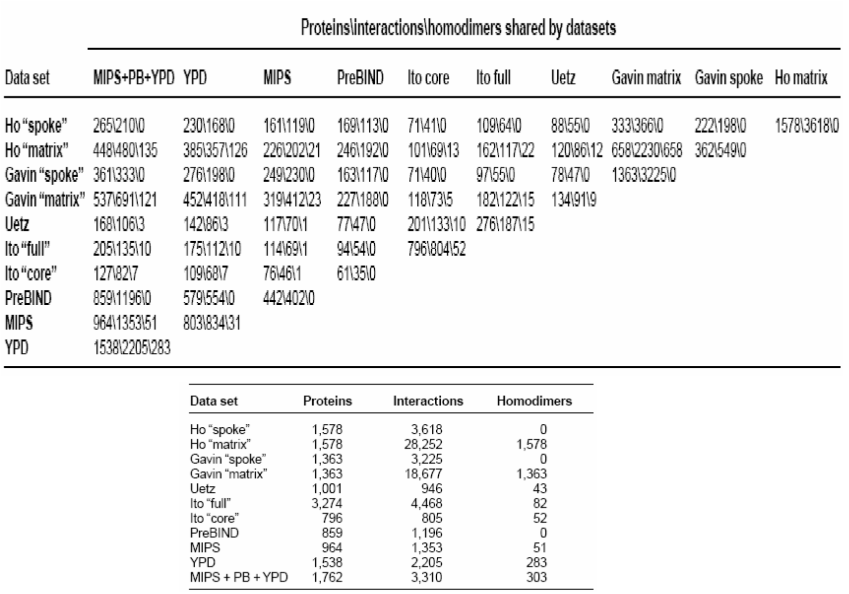
The results show that the reference set is dense along the diagonal, and so is the overlap from high throughput data experiments. HMS-PCI and TAP are dense along the diagonal, but also demonstrate strong density in areas other than the diagonal suggesting a large number of false positives.

1.4 Spoke model vs. Matrix Model

There are two comparison models for translating complexes obtained using coIP methods into binary interactions:

1.4.1 Spoke model

The spoke model assumes that all proteins interact with the tagged protein used for the purification, and do not interact with any other proteins. The latter may lead to false negative results.



	Proteins\interactions\homodimers shared by datasets									
Data set	MIPS+PB+YPD	YPD	MIPS	PreBIND	Ito core	Ito full	Uetz	Gavin matrix	Gavin spoke	Ho matrix
Ho "spoke"	265\210\0	230\168\0	161\119\0	169\113\0	71\41\0	109\64\0	88\55\0	333\366\0	222\198\0	1578\3618\0
Ho "matrix"	448\480\135	385\357\126	226\202\21	246\192\0	101\69\13	162\117\22	120\86\12	658\2230\658	362\549\0	
Gavin "spoke"	361\333\0	276\198\0	249\230\0	163\117\0	71\40\0	97\55\0	78\47\0	1363\3225\0		
Gavin "matrix"	537\691\121	452\418\111	319\412\23	227\188\0	118\73\5	182\122\15	134\91\9			
Uetz	168\106\3	142\86\3	117\70\1	77\47\0	201\133\10	276\187\15				
Ito "full"	205\135\10	175\112\10	114\69\1	94\54\0	796\804\52					
Ito "core"	127\82\7	109\68\7	76\46\1	61\35\0						
PreBIND	859\1196\0	579\554\0	442\402\0							
MIPS	964\1353\51	803\834\31								
YPD	1538\2205\283									

Data set	Proteins	Interactions	Homodimers
Ho "spoke"	1,578	3,618	0
Ho "matrix"	1,578	28,252	1,578
Gavin "spoke"	1,363	3,225	0
Gavin "matrix"	1,363	18,677	1,363
Uetz	1,001	946	43
Ito "full"	3,274	4,468	82
Ito "core"	796	805	52
PreBIND	859	1,196	0
MIPS	964	1,353	51
YPD	1,538	2,205	283
MIPS + PB + YPD	1,762	3,310	303

Figure 5: Source [1]: A comparison among different PPI datasets.

1.4.2 Matrix model

The matrix model assumes that any two proteins within a complex have a pairwise interaction. The matrix model contains all possible true interactions within the data, but may also includes false positives.

Figure 7 shows an analysis of large scale protein interaction data sets [18]. The research used both a matrix model and a spoke model to represent HMS-PCI and TAP data sets, and measure their accuracy. As can be seen in Figure 7, spoke models fit functional data better (functional data will be discussed in Section 1.4). The functional distribution of interactions for the spoke modeled HMS-PCI and TAP data sets resembles more closely the literature and Y2H interactions than the matrix modeled data. Therefore, in spite the fact that information is discarded in the spoke model, it is probably a better model than the matrix model, since it bears closer resemblance to the literature.

1.5 Reliability Assessment By Comparison To Reference

When assessing the quality of interaction data, the coverage and the accuracy are needed to be considered together. A data set of high coverage is not very useful if its accuracy is low, i.e. if it contains a large number of false positives.Similarly, a data set of high accuracy is not very useful if its coverage is low.

Commonly, one estimates accuracy and coverage by comparing to a reference set of known interactions. [18].

Figure 8 shows results from different high throughput experiments such as Y2H, HMS-PCI and TAP, as well as the overlap between 2 and 3 different methods. Each experiment was repeated several times with different parameter choices (for example, matrix model/spoke model; see [18] for a complete list). The various data sets are benchmarked against a reference set of 11,000 known interactions, which are derived from manually annotated protein complexes (such as MIPS). Each point in the graph represents an entire interaction data set and its position specifies coverage and accuracy (on a log-log scale).

Notably the comparison is incomplete since the reference set is still incomplete, and may suffer from biases. However, we can see that there are large differences between the various methods and even within a method when certain parameters are changed. For example, if we consider the spoke model compared to the matrix model (figure 9) we get better accuracy when using the first (6.8% instead of 2% in HMS-PCI, and 37.8% instead of 12.55% in TAP). The coverage however decreases greatly (from 21% to 8.5% in TAP, and from 6.1% to 2.3% in HMS-PCI). These results are not suprising since the spoke model assumes less interactions than the matrix model. If we consider results obtained by 3 methods, we get best accuracy but relatively low coverage. Results obtained by 3 methods are more likely to be true positives, but not all true interactions will be revealed by all three methods.

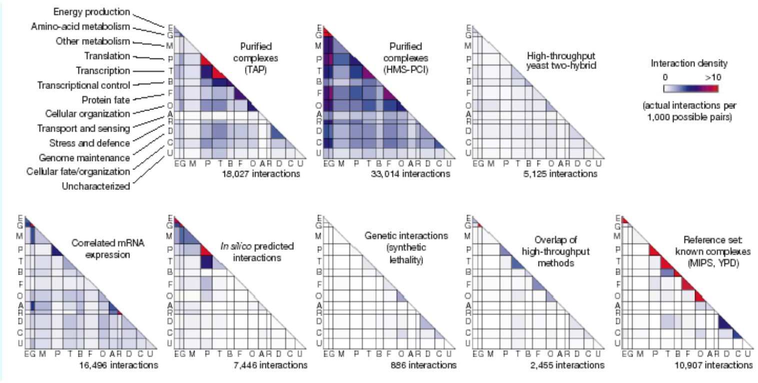


Figure 6: Source [18]: Large scale interaction data and the distribution of interactions according to functional categories. Each data set is represented by a matrix showing the distribution of interactions by color. White color represents a small number of interactions, while red color represents a large number of interactions. Different shades of blue represent middle ranges. Proteins with related functions preferentially interact with each other, and therefore it is expected to see a large number of interactions along the diagonal. Interactions outside the diagonal are likely to be false positives.

1.6 Coverage Bias

Most protein data sets are heavily biased toward proteins of high abundance. This implies that interactions between proteins of low abundance may remain undiscovered. A study by von Mering et al. [18] recorded for each data set and abundance class, the number of interactions having at least one protein in that class. Each interaction (A-B) is therefore counted twice: Once under the abundance class of partner A, and once under the abundance class of partner B. The study shows that genetic approaches such as Y2H are relatively unbiased, while other methods such as HMS-PCI and TAP are highly biased towards high abundance.

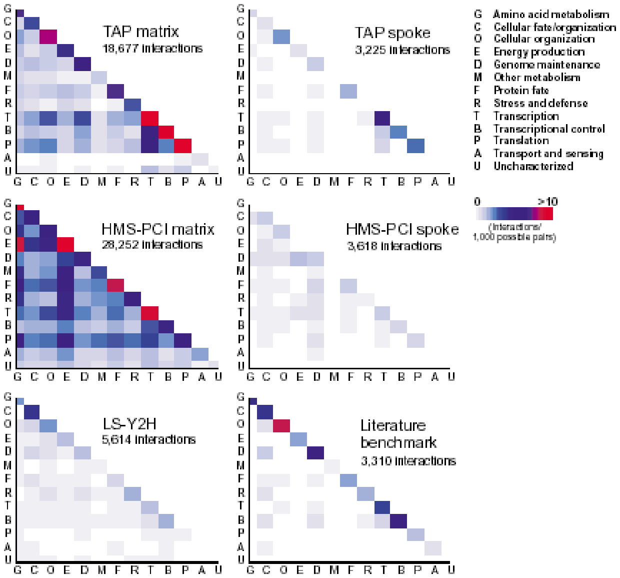


Figure 7: Source [1]: Functional annotation matrices. Distribution of interaction of 6 data sets.

1.7 Estimating the number of Protein-Protein interactions

1.7.1 Sampling overlap

This method, as described in [8] assumes independence of different experiments. Denote the true number of interactions as N_i, the number of interactions obtained using experiment $\sharp 1$ by n_1, the number of interactions obtained using experiment $\sharp 2$ by n_2. Finally denote n_{12} as the number of interactions in the intersection of N_1, N_2.

The probability that an interaction is in N_1 is n_1/n_i. Similarly, the probability that an interaction is in N_2 is n_2/n_i. Hence we obtain: $n_{12} = (n_1 \cdot n_2)/n_i \Leftrightarrow n_i = (n_1 \cdot n_2)/n_{12}$ The average number of interactions per proteins is obtained by calculate the average of N_i over all proteins. Since N_i is the average number of interactions for an individual protein, we should multiply that average by the total number of proteins, then divide by 2 (since each interaction is counted twice - once for each protein). This calculation will give us an estimation of the total number of interactions.

In [8], A. Grigoriev tried to estimate the average number of interactions in yeast. The average N_i he received in yeast was 5. As yeast has 6300 proteins, the total number of interactions estimated by this method is approximately $6300 \cdot 5/2 = 16000$.

1.7.2 Power-law

The integrated PPI network follows a power low node connectivity distribution. By scaling the power low connectivity distribution of a partial data set (that contains 76% of the yeast proteome) to the entire yeast proteome we receive an estimation of ~ 20000 interactions in yeast [8].

1.8 TAP vs. HMS PCI

Bader & Hogue [1] used 115 common baits, and compared the coverage to the reference data set. TAP spoke model revealed 87 out of 628 known interactions, while TAP matrix model revealed 264 out of 4916 known interactions. With HMS-PCI, the spoke model revealed 94 of 875 known interactions, whereas the matrix model revealed 193 out of 7618 known interactions. According to these results TAP model is 32% better at detecting known interactions.

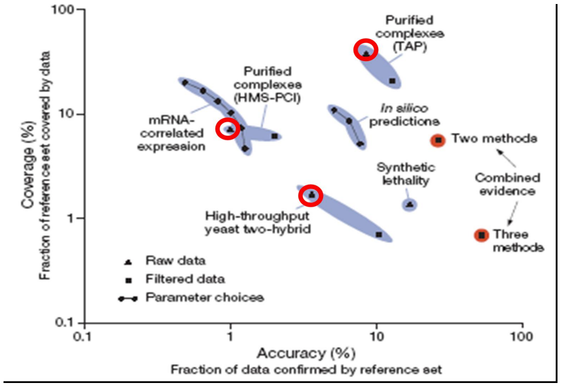
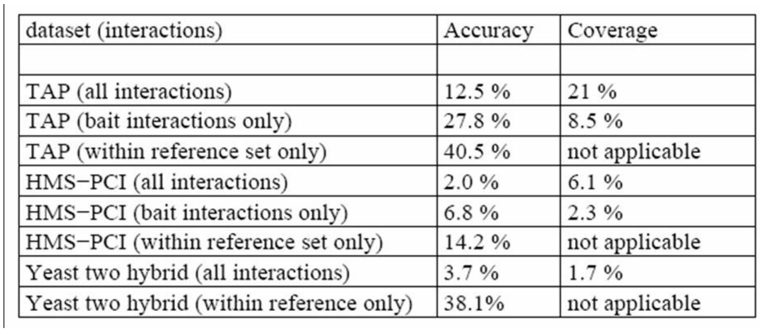


Figure 8: Source [18]: Quantitative comparison of interaction data sets.



dataset (interactions)	Accuracy	Coverage
TAP (all interactions)	12.5 %	21 %
TAP (bait interactions only)	27.8 %	8.5 %
TAP (within reference set only)	40.5 %	not applicable
HMS−PCI (all interactions)	2.0 %	6.1 %
HMS−PCI (bait interactions only)	6.8 %	2.3 %
HMS−PCI (within reference set only)	14.2 %	not applicable
Yeast two hybrid (all interactions)	3.7 %	1.7 %
Yeast two hybrid (within reference only)	38.1%	not applicable

Figure 9: Source [18]: Accuracy and coverage rates for various high scale methods (corresponds to the data in Figure 8).

2 CONFIDENCE ASSIGNMENT SCHEMES

2.1 Validation Methods

In order to validate interaction data a comparison to a reference set can be made. One method is to compare the results to existing databases (such as MIPS - http://mips.gsf.de) - the data is considered to be accurate, but is not exhaustive enough. Another method is comparing the results to small scale experiments - data is also accurate, but scarce and expensive to get. Taking interactions that are supported by two or more experimental techniques is a third method. If an interaction is supported by several methods it is more likely to be true. It will however result in ignoring those true interactions that are only supported by one technique (for example more rare interactions).

As another validation method, one can examine the tendency of interacting proteins to have similar biological annotations. Example annotations include biological process, cellular component, expression pattern and more.

2.2 Correlation with Localization

Figure 11 suggests a positive correlation between interaction and similarity in localization. Interactions revealed in small scale experiments consist of proteins pairs, all of which share the same spatial localization. In large scale experiments the results vary from 40% to 80%. When considering overlapping datasets from high throughput experiments, the localization level is a little higher. These results indicate that a large percentage of the interactions obtained are false positives.

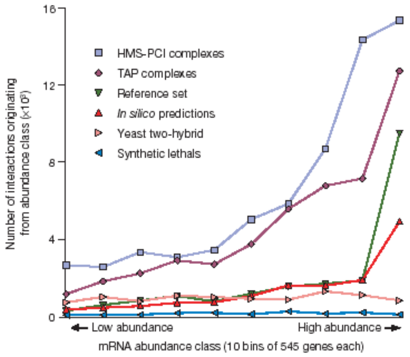


Figure 10: Source [18]: A bias in interaction coverage from mRNA abundance data .

2.3 Correlation with Expression

The study of Deng et al. [5] tried to assess the reliability of protein-protein interaction data. Different data sets were compared to known interactions in MIPS. Deng et al. computed the expression correlation coefficient for every interaction protein pair using cell cycle gene expression data [5]. Figure 12 demonstrates the interactions frequency using a division to bins. The division into the different bins is done according to the expression correlation coefficient. True interactions are more likely to occur within pairs with similar expression. Therefore it is expected that true interactions will have a relatively high correlation coefficient. As we can see in Figure 12, the MIPS graph is shifted more to the right than all the other graphs, which indicates higher values of the correlation coefficient. This result is expected since MIPS data contains only true interactions. HMS-PCI and TAP have lower correlation coefficient values, which indicates those data sets contain false positive interactions.

Figure 13 displays results from two groups:

1. MIPS,DIP, Uets, Ito interaction data: Data sets that contain pairwise physical interactions

2. MIPS Complex, TAP, HMS-PCI: Data sets that contain protein complexes

Among the second group, the MIPS complex has the largest mean expression correlation coefficient which is significantly higher than the results obtained from random pairs is (mean of 0.03). In the first group, MIPS also has the largest mean, while Ito's data has the smallest mean.

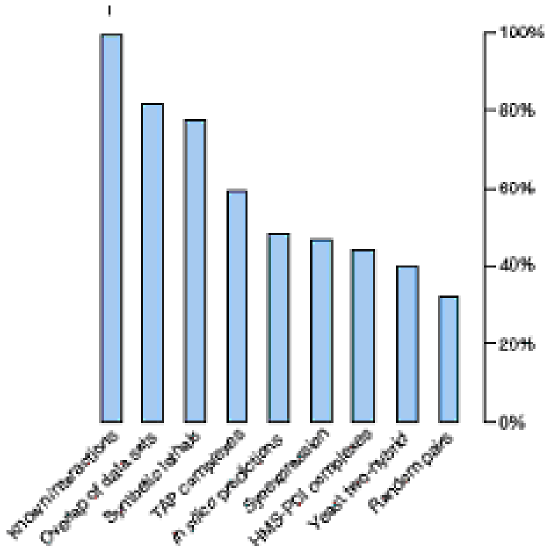


Figure 11: Source [18]: Fraction of interactions in which both partners have the same protein localization

2.4 Correlation with Clustering Coefficient

We expect that interacting proteins will have a large number of joint neighbors, and therefore a relatively high clustering coefficient. The high clustering coefficient of small world networks indicates that neighbors of a given vertex are more likely to have edges between them than would be expected in a random graph.

We define a mutual clustering coefficient C_{vw} for a pair of vertices v and w. We consider 4 different definitions to C_{vw}.

$N(v)$ represents the neighbourhood of a vertex v. $|N(v)|$ represents the size of the neighbourhood. T represents the total number of proteins in the organism.

Jaccard: the Jaccard Index is a statistic used for comparing the similarity and diversity of sample sets. However, the Jaccard Index is inappropriate when one of the two endpoints of a given edge has a large neighborhood. Such situations can be expected in scale-free networks such as that of protein-protein interactions.

$$Jaccard : C_{vw} = \left|N(v) \cap N(w)\right| / \left|N(v) \cup N(w)\right|.$$

Meet/Min: The Meet/Min coefficient removes this bias at the expense of discarding information about the larger neighborhood size. The principal difference is that it is independent of any evidence of an edge between the two nodes measured.

$$Meet/Min : C_{vw} = \left|N(v) \cap N(w)\right| / min(\left|N(v)\right|, \left|N(w)\right|).$$

Geometric: compromise between Meet/Min and Jacard - it does not discards information about large neighbourhood sizes, but it still suffers from some bias, similarly to the Jaccard Index.

$$Geometric : C_{vw} = \left|N(v) \cap N(w)\right|^2 / (\left|N(v)\right| \cdot \left|N(w)\right|).$$

Hypergeometric: The summation in the hypergeometic coefficient can be regarded as a p-value, i.e. the probability of obtaining a number of mutual neighbors between vertices v and w at or above the observed number in random graphs.

$$Hypergeometric : C_{vw} = -log \sum_{i=|N(v) \cap N(w)|}^{min(|N(v)|,|N(w)|)} \frac{\binom{|N(v)|}{i} \cdot \binom{T-|N(v)|}{|N(w)|-i}}{\binom{T}{|N(w)|}}.$$

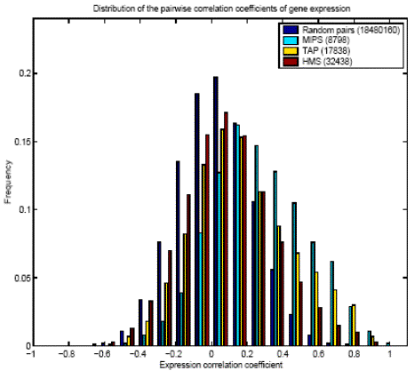
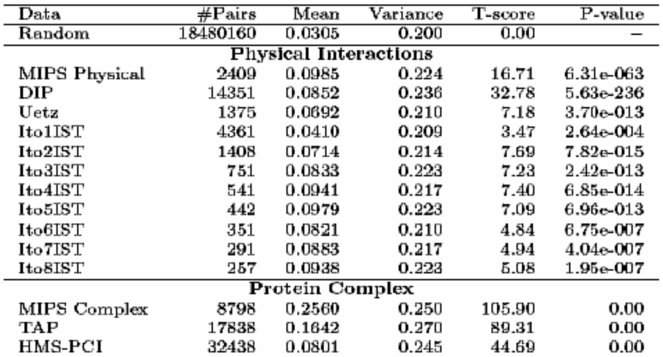


Figure 12: Source [5]: Distribution of the pairwise correlation coefficients of gene expression



Data	#Pairs	Mean	Variance	T-score	P-value
Random	18480160	0.0305	0.200	0.00	–
Physical Interactions					
MIPS Physical	2409	0.0985	0.224	16.71	6.31e-063
DIP	14351	0.0852	0.236	32.78	5.63e-236
Uetz	1375	0.0692	0.210	7.18	3.70e-013
Ito1IST	4361	0.0410	0.209	3.47	2.64e-004
Ito2IST	1408	0.0714	0.214	7.69	7.82e-015
Ito3IST	751	0.0833	0.223	7.23	2.42e-013
Ito4IST	541	0.0941	0.217	7.40	6.85e-014
Ito5IST	442	0.0979	0.223	7.09	6.96e-013
Ito6IST	351	0.0821	0.210	4.84	6.75e-007
Ito7IST	291	0.0883	0.217	4.94	4.04e-007
Ito8IST	257	0.0938	0.223	5.08	1.95e-007
Protein Complex					
MIPS Complex	8798	0.2560	0.250	105.90	0.00
TAP	17838	0.1642	0.270	89.31	0.00
HMS-PCI	32438	0.0801	0.245	44.69	0.00

Figure 13: Source [5]: Statistics of distribution of gene expression correlation coefficients for different PPI data sets

As can be seen from figure 14, the average C_{vw} of interactions found by high confidence studies was higher than the average C_{vw} of interactions found only in Y2H studies. The latter was several orders of magnitude higher than the C_{vw} for pairs of proteins with no evidence of interaction.

2.5 Interologs

Consider two different human protein molecules A and B that interact. Investigating homologous proteins to A and B in another organism ,say A' and B', one might find that A' is homologous to A and B' is homologous to B. The pair A' and B' is called an interolog of A and B if (and only if) A' and B' also interact.

Interologs are interactions between species which indicate also likely to interact Therefore, interaction maps from one species might be useful in predicting interactions in other species. These conserved interactions between species are called interologs.

Matthews [11] started from a large number of published Y2H interactions between yeast proteins, and searched for candidate interologs in a certain type of worm (C. elegans). Between 16% -31% of protein interactions in the worm could be detected in a Y2H experiment, suggesting that these interactions are indeed conserved. These remarkable results are in spite the fact that the yeast and worm are 900 million years apart.

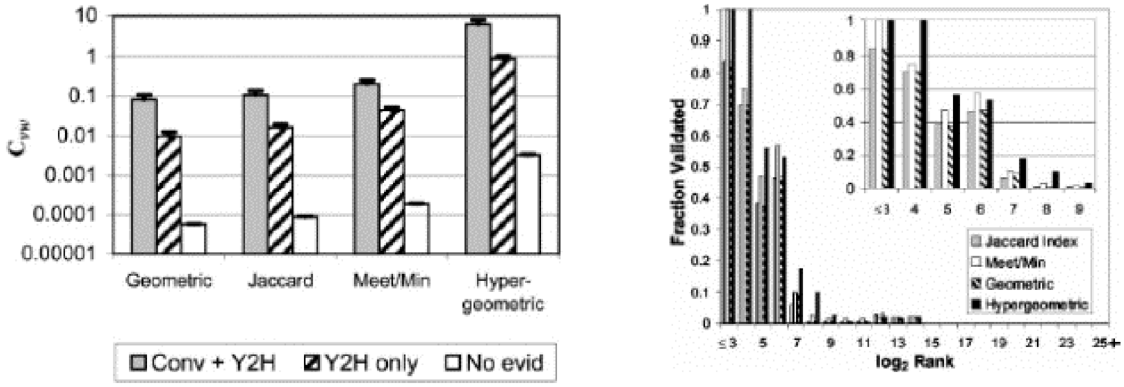


Figure 14: Source [7]: Left: Average mutual clustering coefficient (C_{vw}) of interactions found in both conventional studies and Y2H studies (CONV: Small scale experiments, No Evid: No evidence of interaction) Right: Fraction validated by C_{vw}. (all pairs of proteins are ranked by C_{vw} and logarithmically binned)

2.6 Comprehensive Correlations

2.6.1 ROC Curve

Experiments sch as Y2H and coIP provide us information on the status of interaction among pairs of proteins. These are four categories in which the experiments results fall:

1. True Positive (TP): An interaction between proteins is predicted, and in reality the proteins also interact.
2. True Negative (TN): An interaction between proteins is not predicted, and in reality the proteins indeed don't interact.

   Incorrect Classifications:

3. False Positive (FP): An interaction between proteins is predicted, but in reality the proteins don't interact.
4. False Negative (FN): An interaction between proteins is not predicted, but in reality the proteins do interact.

One can consider incorrect classification as a false result, without regarding the type of error (false positive or a false negative) . However, in real life we may want to differentiate between the two types of errors. For example, if accuracy is utmost important, we may care very little for false negative errors, but refuse to accept false positive errors. We define Sensitivity as the probability that an interaction will be revealed by an experiment, given that the interaction occurs in reality. Specificity is the probability that an interaction will not be assumed by the experiment, given that it does not occur in reality. Therefore the sensitivity is given by True Positives Rate- TPR=TP/(TP+FN), and the specificity is given by 1 minus False positives Rate (FPR)- FPR = FP/(FP+TN).

A ROC curve graphically displays the performance of a classification method for different costs i.e. sensitivity vs. specificity. It consists of a set of points, each computed for a different setting of the cost, connected by lines. For each point the vertical coordinate is a true positive rate (or sensitivity), while the horizontal coordinate is a false positive rate (or 1-specificity). The ROC curve for an optimal classifier will be as close as possible to the upper left corner of the chart (where we have the highest number of true positives and the smallest number of false positives).

The ROC curve demonstrated in Figure 15 is illustrated by: TPR and FPR - The positives are MIPS complexes, and the negatives are pairs from different compartments. As can been seen from the graphs in Figure 15, INT, MIP, GOF and COR give the best results - their curve is the closest to the upper left corner.

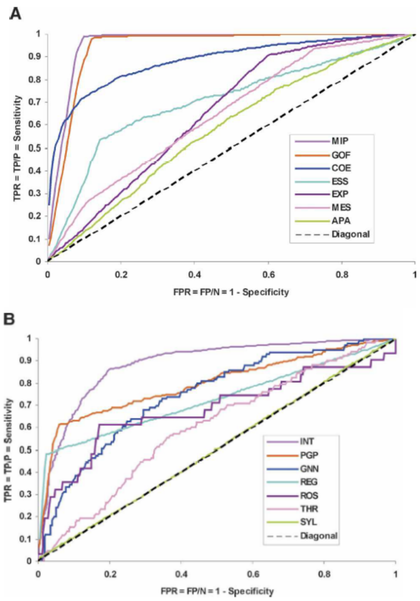


Figure 15: Source [10]: (A+B): Predictive power of individual features illustrated by ROC curves INT - Interolog mapping, MIP - Mips functional similarity, GOF - Go functional similarity), COR - mRNA co expression

2.7 ASSIGNING RELIABILITIES TO DATA SETS

The reliability of a data set of protein-protein interactions is defined as the fraction of real protein interactions over all reported interactions.

2.7.1 Expression Profile Reliability

The EPR (expression profile reliability) index [4] estimates the fraction of biologically relevant protein interactions detected in a high throughput screen. It does so by comparing the RNA expression profiles for the proteins whose interactions are found in the screen with expression profiles for known interacting and non interacting pairs of proteins.

The reference data used by Deane et al is taken from DIP [16] and includes 2000 interactions that were obtained from small scale experiments. All the pairs of proteins that are not in DIP were used as negative data.

The overall distribution of expression distances obtained for an experimental set is composed of the distribution for interacting proteins with probability $alpha_{EPR}$ and the distribution for non-interacting proteins with probability $(1 - alpha_{EPR})$.

We denote: d() as the Eucleadian distance between expression patterns and ρ_i and ρ_n as the expression distance probability distribution for interacting and non-interacting protein pairs, respectively. The index $alpha_{EPR}$ is defined as the percentage of true positives. Then, the resulting, overall distribution of expression distances obtained for an experimental set is described by:

$$\rho_{\exp}(d_{AB}^2) = \alpha_{EPR} \cdot \rho_i(d_{AB}^2) + (1 - \alpha_{EPR}) \cdot \rho_n(d_{AB}^2)$$

α_{EPR} can be found by linear least-sqaure fit method.

Figure 17 shows the EPR index as calculated for several subsets of DIP. The α_{EPR} parameter was evaluated by using linear least-squares fit. As can be seen, the CORE method is found to be the most accurate by this method, with an EPR index of 0.92.

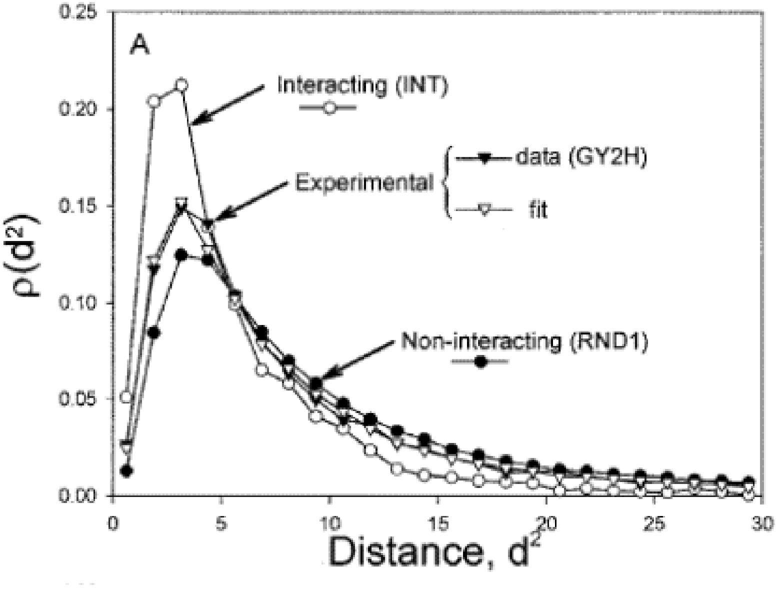


Figure 16: Source [4]: evaluation of genome wide Y2H interaction data in DIP using the EPR index.

2.7.2 Noise estimation

Deng et al [5] wish to estimate noise via gene expression. We denote O() as the distribution of the correlation coefficients of the gene expression profiles for the given set's alleged interactions. T() is the distribution for the true interacting pairs (true positives) and R() is the distribution for the non interacting pairs (false positives).

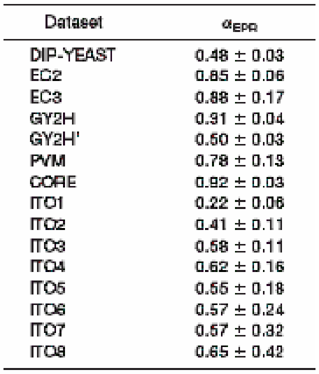
Deng et al. splits the values of the correlation coefficients into k bins. Let n_k be the number of observed interaction pairs in the k_{th} bin. Let p_k and q_k be the fraction of real interacting pairs and random pairs in the k_{th} bin, respectively. Let α is the reliability of a data set of putative interactions.

The likelihood function is defined as:

$$L(\alpha) = \prod_k (\alpha p_k + (1 - \alpha) q_k)^{n_k}$$

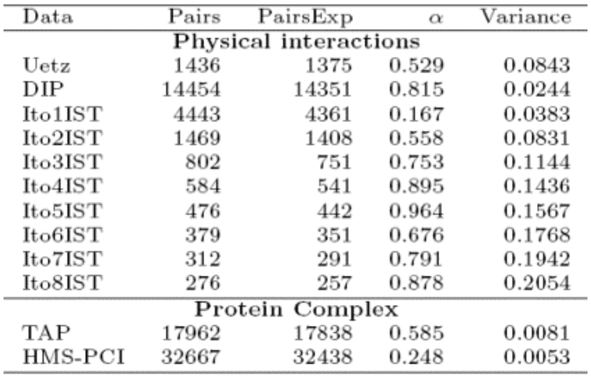
L is a convex function, therefore, α can be estimated by maximizing $L(\alpha)$ using classical gradient algorithms, described in section 2.8.

Among the first group (physical interactions), Ito5IST is the most reliable Among the second group (protein complexes), TAP is more reliable than HMS-PCI. These experiments were conducted in 2003, and that is probably one of the reasons why DIP is more reliable according to Deng et al.'s estimation.



Dataset	α_{EPR}
DIP-YEAST	0.48 ± 0.03
EC2	0.85 ± 0.06
EC3	0.88 ± 0.17
GY2H	0.31 ± 0.04
GY2H'	0.50 ± 0.03
PVM	0.78 ± 0.13
CORE	0.92 ± 0.03
ITO1	0.22 ± 0.06
ITO2	0.41 ± 0.11
ITO3	0.58 ± 0.11
ITO4	0.62 ± 0.16
ITO5	0.55 ± 0.18
ITO6	0.57 ± 0.24
ITO7	0.57 ± 0.32
ITO8	0.65 ± 0.42

Figure 17: Source [4]: EPR index: α_{EPR} calculated for several subsets of DIP. DIP - Database of Y2H interacting proteins, ECX - Results obtained from X (or more) different experiments, GY2H - Y2H with ITO1 results excluded, PVM - interacting paralogs, CORE: PVM + INT + EC2, ITOX: Results obtained by X ito experiments



Data	Pairs	PairsExp	α	Variance
Physical interactions				
Uetz	1436	1375	0.529	0.0843
DIP	14454	14351	0.815	0.0244
Ito1IST	4443	4361	0.167	0.0383
Ito2IST	1469	1408	0.558	0.0831
Ito3IST	802	751	0.753	0.1144
Ito4IST	584	541	0.895	0.1436
Ito5IST	476	442	0.964	0.1567
Ito6IST	379	351	0.676	0.1768
Ito7IST	312	291	0.791	0.1942
Ito8IST	276	257	0.878	0.2054
Protein Complex				
TAP	17962	17838	0.585	0.0081
HMS-PCI	32667	32438	0.248	0.0053

Figure 18: Source [5]: Statistics of distributions of gene expression correlation coefficients for different protein-protein interactions data sets.

2.8 ASSIGNING RELIABILITIES TO INDIVIDUAL INTERACTIONS

2.8.1 Assignment Procedure

Each interaction has a set of attributes associated with it, such as experiment type (small scale, Y2H, coIP, other large scale experiment), protein expression, GO annotation etc. We wish to predict according to this set of attributes if the interaction occurs or not. We need to select the following: a set of attributes that will be used to predict the interaction (classification features). Out of all the interactions we will take a subset, along with its sets of attributes and the prediction, as training data. Finally we will need a classification algorithm which will allow us to predict whether or not the interaction will occur.

2.8.2 Training Data: Y2H vs. coIP

A crucial step in analyzing interaction data is separating the subset of credible interactions from the background noise. Bader at el [2] selected interactions for positive and negative sets by comparing protein networks constructed from pubished Y2H and copIP data (in matrix model). Instead of considering only interactions found in both data sets, Bader examined a contingency table in which pairs of proteins were binned according to the distances between the paired proteins in each of the two networks.

The color of the graph in Figure 19 is determined by the number of observed interactions divided by the number of expected interactions (in random). The redder the square, it is more likely we have a significant result. The distance

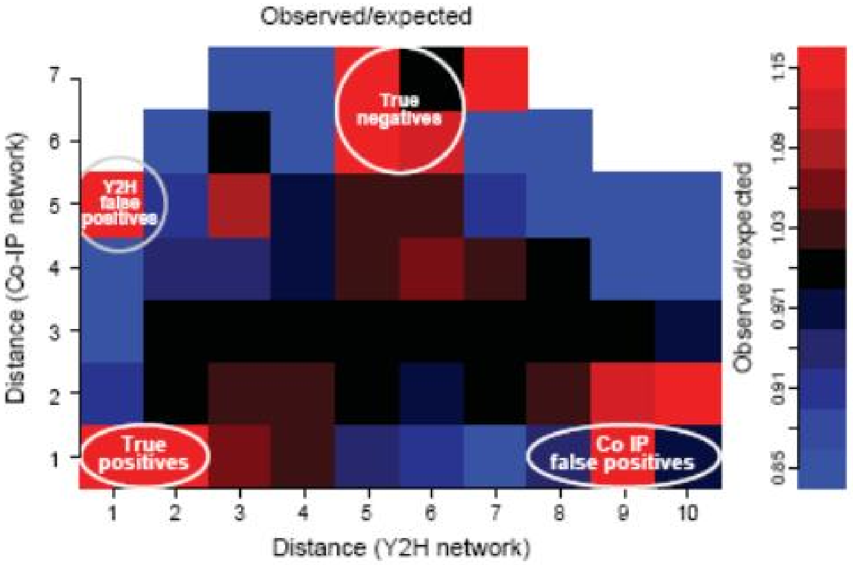


Figure 19: Source [2]: Network cross-comparison: pairs of proteins binned according to their shortest path in networks generated from Y2H and COIP data.

in a graph is defined as the number of edges in coIP the shortest path from one node to another.

The positive training set was selected as interactions between proteins that were 1-2 edges apart in the Y2H network (marked in Figure 19 as True Positives). The negative training set was selected as interactions in one training set between proteins whose distance in the other training set was larger than the median separation for random pairs of proteins (marked as True Negatives). Note that sets located in the lower right corner and upper left corner are probably false positives - sets that are supported by one experiment only.

2.8.3 Model Variables

Different schemes will consider different input attributes in order to predict interactions. In their experiment Sharan et al. [14] used binary variables denoting whether or not interaction was predicted under that technique: Y2H, COIP, small scale, large scale (other than Y2H, CoIP). Bader [2] considers indicator for high quality data, clustering related measures (how big the cluster is, how close the proteins are), number of complexes in which the interaction does not involve the bait (in coIP/Matrix model) Another scheme by Bader [2] considers also plus GO annotation, expression correlation and existence of genetic interactions as input attributes.

2.8.4 The logistic function

The logistic function, $1/(1+e^{-x})$, is widely used in classification problems.

It attains values between 0 and 1. Its first derivative is especially high around 0, thus the change in that area will be rapid. The further we get from the 0 point, the smaller the derivative is, and hence the change.

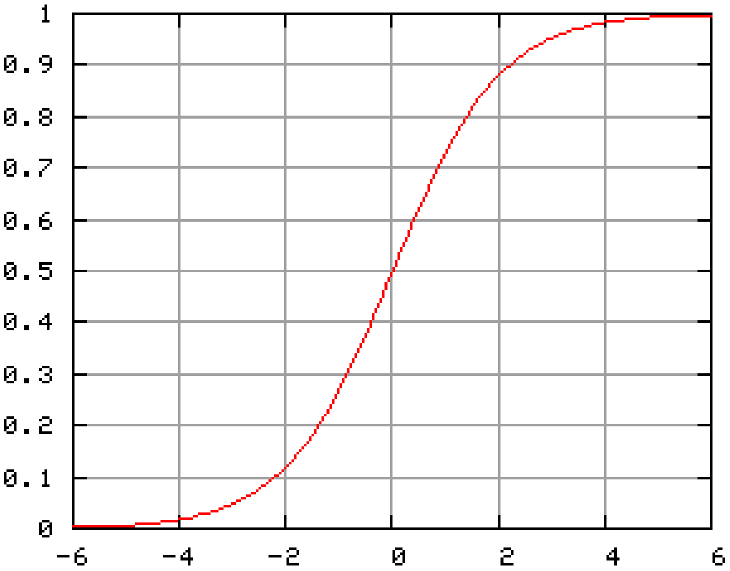


Figure 20: The logistic function.

2.8.5 The logistic function and binary classification [12]

Consider the problem of assigning a binary label Y to a vector X with n attributes.

By applying Bayes rule, we get:

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)} = \frac{1}{1+\exp\{-\log\frac{P(X|Y=1)}{P(X|Y=0)} - \log\frac{P(Y=1)}{P(Y=0)}\}}$$

For data sets with multivariate Gaussians with identical covariance matrix (i.e., for every $i = 1,..,n$ and $j = 0,1$, $P(X_i|Y_j)$ is a Gaussian), there exist a vector w and a constant b, such that:

$$P(Y=1|X) = logit(w^t X + b)$$

which is a linear expression inside the logistic function, whose behavior is described in Figure 20.

2.8.6 Logistic regression [12]

Discriminative model: we want to estimate the vector w, in the logistic function:

$$P(Y=1|x) = logit(w^t x + b) = \frac{1}{1+\exp(-w^t x - b)}$$

This can be achieved by choosing a vector w that maximizes the conditional data likelihood. The conditional data likelihood is the probability of the observed Y values in the training data, conditioned on the corresponding X values. In other words,

$$w \leftarrow \max_w \prod_l P(Y^l|X^l, w)$$

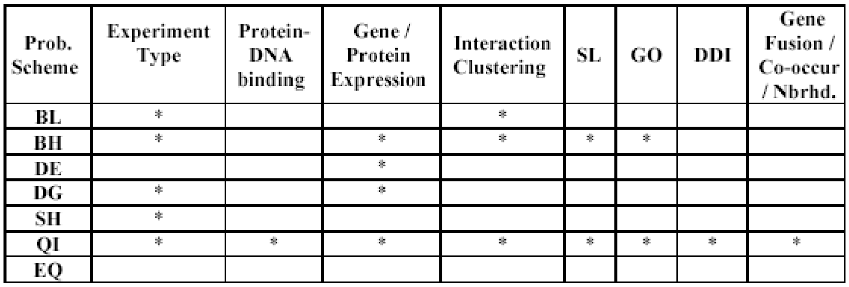
where X^l and Y^l are the observed values of X and Y in the l_{th} training example, respectively. Equivalently, the following likelihood function can be maximized:

$$L = \sum_{x=1} \log[logit(w^t x + b)] + \sum_{x=0} \log[1 - logit(w^t x + b)]$$

L can be shown to be concave, hence it can be maximized using a gradient ascent process. In each interaction we move in the direction of the gradient until we reach the global maximum. The gradient is given by:

$$\begin{aligned} \mu(x) &= logit(w^t x + b) \\ \Delta_w(L) &= \sum_{x=1}[1 - \mu(x)]x + \sum_{x=o}\mu(x)x \end{aligned}$$

3 Comparison Between PPI Reliability Estimation Schemes (Southern et al.)



Prob. Scheme	Experiment Type	Protein-DNA binding	Gene / Protein Expression	Interaction Clustering	SL	GO	DDI	Gene Fusion / Co-occur / Nbrhd.
BL	*			*				
BH	*		*	*	*	*		
DE			*					
DG	*		*					
SH	*							
QI	*	*	*	*	*	*	*	*
EQ								

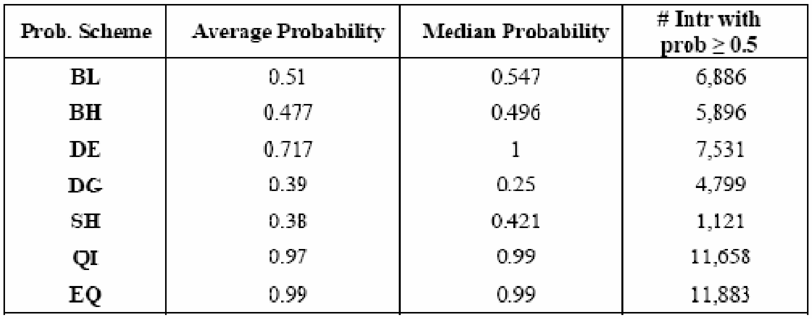
Figure 21: Source [15]: summary of input attributes for the different probability schemes BL/BH/SH: logistic regression, DE: Deane, DG: Deng, EQ: Equal (all observed interactions are considered to be equally true), QI.

High throughput methods like Y2H might suffer from high rates of false positive detection. To combat these errors, many methods have been developed which associate confidence scores with each interaction. Suthram et al [15] perform an analysis and performance assessment of these different methods, using the fact that interacting proteins have similar biological attributes such as function, expression, and evolutionary conservation. In addition a new measure called signal to noise ratio is introducted.

A set of 11,883 interactions common to all the tested schemes were taken. 6 measures were used to assess the accuracy of each interaction probability scheme. A summary of all the attributes used as input in different probability schemes can be seen in Figure 21.

3.1 Global properties

Basic statistics on the compared schemes are given in Figure 22:



Prob. Scheme	Average Probability	Median Probability	# Intr with prob ≥ 0.5
BL	0.51	0.547	6,886
BH	0.477	0.496	5,896
DE	0.717	1	7,531
DG	0.39	0.25	4,799
SH	0.38	0.421	1,121
QI	0.97	0.99	11,658
EQ	0.99	0.99	11,883

Figure 22: Source [15]: Comparison of global properties of different probability assignment schemes

QI and EQ both have the largest number of interactions with probability above 0.5. These methods have a very high average and median probability. SH has the lowest values.

3.2 Function and Expression Correlation

The number of proteins assigned to the term was evaluated by the deepest common GO term assignment (deepest common ancestor) shared between a pair of proteins that interacts. Spearman, mutual information and weighted average were calculated in order to assess the relationship between the size of the deepest common GO term and interaction probabilities.

3.3 Correlation

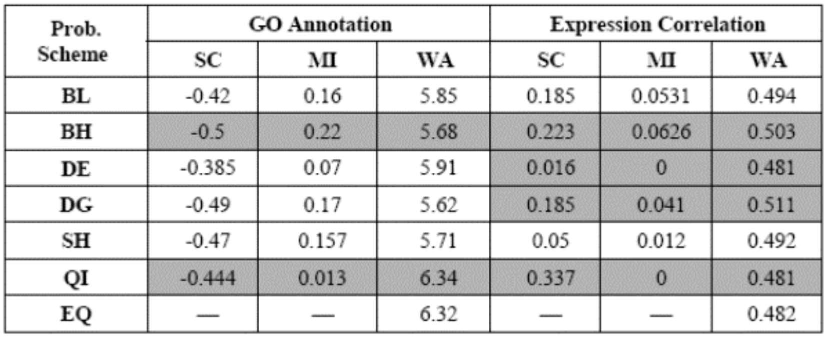
In order to calculate the expression correlation, Suthram et al. used yeast expression data for ~ 700 conditions that were obtained from a database (SMD). Spearman, mutual information and weighted average between the expression correlation coefficients of interacting proteins and their corresponding probability assignment in different scheme were calculated.

WA: Weighted Average: The average which takes into account the proportional relevance of each component, instead of treating them equally. MI: Mutual information: A quantity that measures the mutual dependence of the two random variables.

$$I(X;Y) = D(p(x,y)||p(x)p(y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

SC: Spearman Correlation = Assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables

$$1 - \frac{6\sum_{i=1}^{n}(rank(X_i) - rank(Y_i))^2}{n(n^2-1)}$$



Prob. Scheme	GO Annotation			Expression Correlation		
	SC	MI	WA	SC	MI	WA
BL	-0.42	0.16	5.85	0.185	0.0531	0.494
BH	-0.5	0.22	5.68	0.223	0.0626	0.503
DE	-0.385	0.07	5.91	0.016	0	0.481
DG	-0.49	0.17	5.62	0.185	0.041	0.511
SH	-0.47	0.157	5.71	0.05	0.012	0.492
QI	-0.444	0.013	6.34	0.337	0	0.481
EQ	—	—	6.32	—	—	0.482

Figure 23: Source [15]: Association of interaction probabilities with the size of the deepest common ancestor in the Gene Ontology assignments and mRNA expression correlation.

The results of the comparison are summerized in Figure 23.

GO Annotation: BH has the maximum correlation with the GO term assignments. The results are not surprising since GO assumptions are taken as input to probability calculation in this scheme. DG also shows good results over the 3 measurements.

Expression Correlation: Bh, Bi, QI, DG all showed a significant association between expression correlations and probabilities. Again, this result is expected since these schemes since all those methods (except for BI) take the expression correlation as input to the probability calculation in the scheme.

3.4 Interologs (Interaction Conservation)

Presence of conserved interactions across species is believed to be associated with biologically meaningful interactions. However, since most species' interaction networks are still incomplete, it is important not to skew the results of this analysis due to false-negatives. In a workshop by Suthram et al. [15], an interaction was considered conserved if the orthologs of the interacting proteins are also interacting. Moreover, Putative orthologs were assigned based on sequence similarity computed using BLAST. The weighted average between probability assignment for each yeast interaction was evaluated as well as the number of conserved interactions across worm and fly. This was repeated for different BLAST E-value thresholds for homology assignments.

3.5 Conservation Coherency

Interacting proteins show a clear preference to be conserved as a pair. For every pair of interacting proteins the conservation rate discrepancy score is calculated as the absolute value of the difference between the evolutionary rates of the two corresponding genes. A low score indicates highly coherent conservation rates.

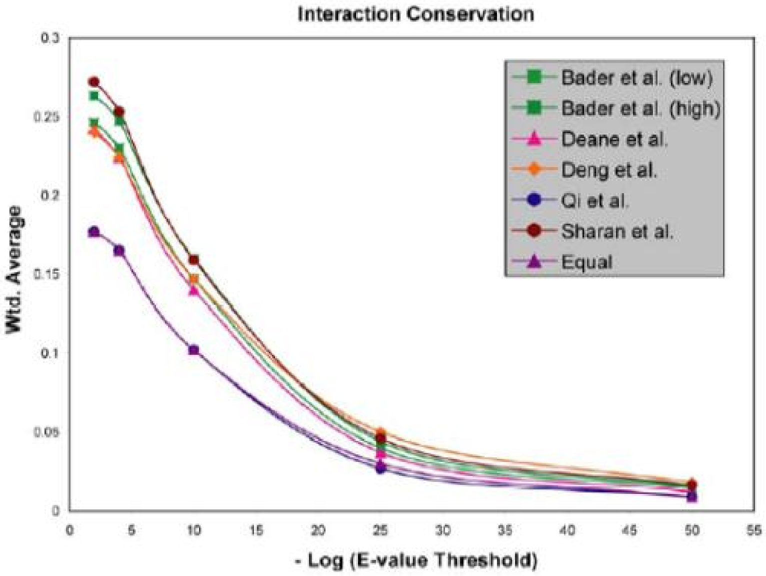


Figure 24: Source [15]: Correlation of number of conserved interactions and probability assignments to interactions

For all the probability assignments schemes Suthram et al. obtain a statistically significant negative correlation between the conservation rate discrepancy cores and the corresponding probabilities (Figure 24), indicating that proteins with high probability interactions tend to have similar conservation rates. As seen in Figure 24, The highest correlation was obtained by DG (Deng).

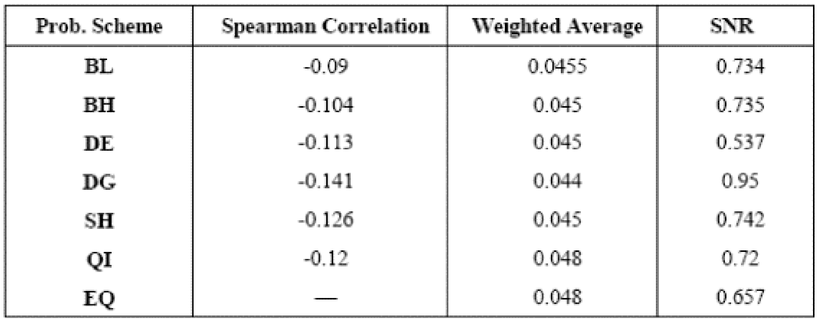
3.6 SNR - Signal To Noise Ration of Protein Complexes

Most cellular processes involve proteins that act together in complexes. We apply a method to find such complexes in yeast , then evaluate the SNR (Signal to noise ratio) is order to assess data quality. To compute SNR, a search for dense interaction complexes is initiated from each node of the graph (=protein) and the highest scoring complex from is taken. This yields a distribution of complex scores over all nodes in the graph. We also calculate scores from random graphs with the same degree distribution as the original network.

SNR is then calculated by the following formula:

$$SNR = \log_{10}\frac{rms(original\ complex\ scores)}{rms(random\ complex\ scores)},\quad where\ rms(x) = \sqrt{\frac{1}{M}\sum_{i=1}^{M} x_i^2}$$

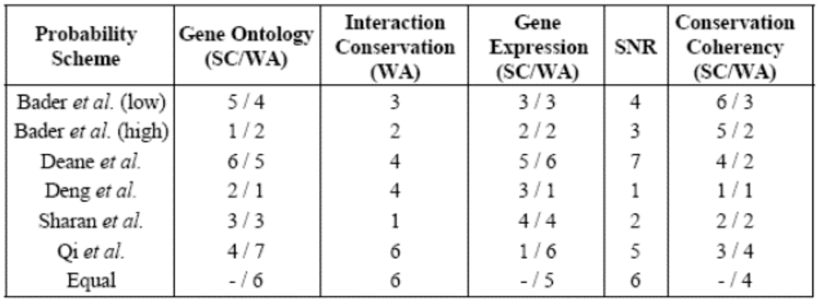
(rms = root mean square)



Prob. Scheme	Spearman Correlation	Weighted Average	SNR
BL	-0.09	0.0455	0.734
BH	-0.104	0.045	0.735
DE	-0.113	0.045	0.537
DG	-0.141	0.044	0.95
SH	-0.126	0.045	0.742
QI	-0.12	0.048	0.72
EQ	—	0.048	0.657

Figure 25: Source [15]: Associations of conservation rate coherency scores and SNR with interaction probabilities

DE and EQ probabilities have low SNR, while SH and DG have the highest SNR values.



Probability Scheme	Gene Ontology (SC/WA)	Interaction Conservation (WA)	Gene Expression (SC/WA)	SNR	Conservation Coherency (SC/WA)
Bader *et al.* (low)	5 / 4	3	3 / 3	4	6 / 3
Bader *et al.* (high)	1 / 2	2	2 / 2	3	5 / 2
Deane *et al.*	6 / 5	4	5 / 6	7	4 / 2
Deng *et al.*	2 / 1	4	3 / 1	1	1 / 1
Sharan *et al.*	3 / 3	1	4 / 4	2	2 / 2
Qi *et al.*	4 / 7	6	1 / 6	5	3 / 4
Equal	- / 6	6	- / 5	6	- / 4

Figure 26: Source [15]: Ranking of the probability schemes in the five measures used for assessing their quality (schemes with rank 1 performs the best).

Figure 26 gives the relative ranking of aforementioned schemes over the five measures used to assess their reliability. EQ almost always ranks the lowest, suggesting that a having probability scheme is better than consider all observed interactions to be true. Deane is the only scheme that assigns reliabilities to a set of interactions, rather than individual interactions thus it performs poorly compared to the other interaction schemes. As can be seen Deng, Sharan and Bader have the best average ranks.

References

[1] G.D. Bader and C.W.V. Hogue. Analyzing protein-protein interaction data obtained from different sources. *Nature Biotech*, 20:991–997, 2002.

[2] J.S. Bader, A. Chaudhuri, M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1):78–85, 2004.

[3] V. Colizza, A. Flamminia, A. Maritanb, and A. Vespignani. Characterization and modeling of protein-protein interaction networks. *Physica A*, 352:1–27, 2005.

[4] C.M. Deane, L. Salwinski, L. Xenarios, and D. Eisenberg. Protein interactions : Two methods for assessment of the reliability of high throughput observations. *Molecular and Cellular Proteomics*, 1:349–356, 2002.

[5] M. Deng, F. Sun, and T. Chen. Assessment of the reliability of ppis and protein function prediction. *PSB*, pages 140–151, 2003.

[6] A.C. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141, 2002.

[7] D.S. Goldberg and F.P. Roth. Assessing experimentally derived interactions in a small world. *PNAS*, 100(8):4372–4376, 2003.

[8] A Grigoriev. On the number of ppis in the yeast proteome. *Nucliec Acides Research*, 31(14):4157–4161, 2003.

[9] T. Ito et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, 98:4569, 2001.

[10] L.J. Lu et al. Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, 15:945–953, 2005.

[11] L.R. Matthews, J. Vaglio, P. ans Reboul, H. Ge, B.P. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence-based searches for conserved ppis or ”interologs”. *Genome Research*, 11:2120–2126, 2001.

[12] T.M. Mitchell. *Generative and discriminative classifiers: naive bayes and logistic regression*. McGraw Hill, http://www.cs.cmu.edu/ tom/mlbook/NBayesLogReg.pdf, 2005.

[13] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17:1030, 1999.

[14] R. Sharan et al. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A*, 102(6):1974–1979, 2005.

[15] S. Suthram, T. Shlomi, E. Ruppin, R. Sharan, and T. Idekrer. Comparison of protein-protein interaction confidence assignment schemes. *BMC Bioinformatics*, 7:360, 2006.

[16] UCLA. database of y2h interacting proteins, 1999-2004. http://dip.doe-mbi.ucla.edu/.

[17] P. Uetz et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623, 2000.

[18] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.